DOCUMENT RESUME

ED 340 772 TM 018 028

AUTHOR Burke, Paul
TITLE You Can Lead Adolescents to a Test But You Can't Make Them Try. Final Report.
SPONS AGENCY Congress of the U.S., Washington, D.C. Office of Technology Assessment.
PUB DATE 14 Aug 91
CONTRACT OTA-H3-6110.0
NOTE 42p.; Contractor report prepared for the Office of Technology Assessment titled "Testing in American Schools: Asking the Right Questions." For related document, see TM 018 025.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment; Estimation (Mathematics); *Evaluation Utilization; Information Dissemination; Mass Media Role; *National Programs; National Surveys; *Scoring; Secondary Education; Standardized Tests; *Student Motivation; Testing Problems; *Testing Programs
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

The strengths and weaknesses of the National Assessment of Educational Progress (NAEP) are discussed. The NAEP estimates the number of students who are more likely to do certain problems correctly than other students. NAEP reports the numbers, briefly describes the problems, and says that more students need to do these problems correctly. The press largely reports this news as presented. Reporters are not usually investigators, and the news media, which are not refereed journals, do not conduct their own reviews to see if NAEP reports are correct. More lead time for reporters and having each NAEP report give a clear summary of the limitations of the data would help improve coverage and the dissemination of NAEP information. However, the limitations of NAEP data are significant, particularly those caused by lack of student motivation to do well on the test and difficulty in describing the knowledge shown by students. These characteristics weaken policymakers' ability to draw conclusions from NAEP results. An appendix provides 12 graphs and 1 table that illustrate how student scores are calculated and how NAEP estimates the likely proficiency of students. (Author/SLD)

YOU CAN LEAD ADOLESCENTS TO A TEST

BUT YOU CAN'T MAKE THEM TRY

Paul Burke, Final Report, August 14, 1991

YOU CAN LEAD ADOLESCENTS TO A TEST

BUT YOU CAN'T MAKE THEM TRY

Paul Burke, Final Report, August 14, 1991

Office of Technology Assessment Contract # H3-6110.0

Page Contents

The opinions in this paper are the author's and are not necessarily those of the Office of Technology Assessment or the US Government.

ABSTRACT

The National Assessment of Educational Progress (NAEP) estimates the number of students who are more likely to do certain problems right than other students. NAEP reports the numbers, briefly describes the problems and says more students need to do these problems right. The press largely reports the news as presented. Reporters are not usually investigators, and news media are not refereed journals, so they do not conduct their own reviews to see if NAEP reports are right. More lead time for reporters, and having each NAEP report give a clear summary of the limitations of the data, would help improve coverage. However the limitations of the data are significant: especially lack of student motivation, and difficulty in describing the knowledge shown by students. These weaken our ability to draw conclusions from NAEP results.

NAEP PROCEDURES TO ESTIMATE STUDENT SKILLS

Number of Questions

In NAEP each student has a few questions on each of several topic areas. For example in the 1990 8th grade math tests, there were 5 topic areas, and an average of 12 questions per student per topic. Each student received one of 7 different test booklets, each of which covered all 5 topics. The booklets gave the students varying numbers of questions on the topics, as shown in the following table:

	Questions per Student	
	Average	Distribution
Numbers and Operations	20	13, 18, 19, 20, 21, 23, 24
Measurement	9	7, 8, 8, 8, 9, 10, 13
Data Analysis, Statistics and Probability	8	6, 7, 8, 8, 9, 9, 10
Geometry	11	9, 10, 10, 11, 11, 12, 15
Algebra and Functions	11	9, 10, 10, 11, 11, 12, 12

Source: Technical Report, pp. 22, 140, 247-52 [1]

Some questions are easy, some moderate, some hard. Depending on the pattern of which questions a

student gets right, NAEP estimates how likely it is this student is very good, poor, or middling [more fully described, with references, in the appendix to this paper]. A student who answers all the problems right is likely to be a good student (though a perfect score might happen by guessing, or by the luck of knowing these specific problems, so NAEP recognizes there is some chance this student is only middling or poor). A student who misses some problems is considered by NAEP as likely to be a middling student. However NAEP recognizes she might be a lucky poor student or an unlucky excellent student who misses problems for a host of reasons (has no incentive to try on this test, hasn't been taught these topics, works quickly and makes careless mistakes, works carefully on easy or interesting problems scattered around the test and doesn't finish, etc.). Thus when a reader sees low scores reported on a NAEP test (or most other tests), the reader must consider how likely it is that these scores measure knowledge of the topic area, versus motivation, luck, speed, etc.

Masters [2] criticizes tests that confound ability in a field with speed or with whether the student has been taught the topic. Hambleton [3] suggests the need for several independent variables measuring these aspects. NAEP believes that its interpretations are correct, even though several dimensions are treated as one in the calculations [4].

NAEP ignores problems after the last one the student does in each 15 minute block of questions, but marks as wrong most of the questions that are skipped over without being answered: skipped questions get counted as right only about 1/x of the time, where x is the number of answer categories in the question [5]. Students however are not told that they should do the problems in order, or that they do not need to try to finish the test, so they may skip around or guess at hard questions at the end of the test, lowering the estimates of their skills.

The scale from poor to good on these tests is called "proficiency" or "grasp" in NAEP [6], and "ability" in most of the literature [7]. NAEP's terms are better ones for them, since NAEP tests measure how much the students have been taught, as well as innate ability. Some might say that an even better term would be "display," since the test measures what the student is willing to display under the test conditions. A student may be more proficient than he or she displays on this test.

Motivation to Do Well on the Test

Some evidence on this motivation factor is available from students' performance on tests required for high school graduation in certain states. We can compare the results when these tests were required for high school graduation, to field tests in previous years when there were no penalties for poor scores.

The following data show that when a serious incentive is present (high school graduation) scores are usually higher. The exceptions are English and composition in Louisiana, and reading in Montgomery County, in all of which the scores were fairly high already in the field tests. The differences seem especially pronounced for blacks and hispanics, to the small extent data are available. The change in incentives is combined with a change in student preparation, which will be discussed more below.

Passing Grades as Percent of Students Taking the Test the First Time

	1991	90	89	88	87	86	85	84	83	82	81	80	79	78	77
Louisiana, the first two lines are grade 11, others are grade 10															
Science	89	**87**	**71***	**69***											
Social Studies	88	**89**	**77***	**70***											
Mathematics	83	82	**77**	**71***											
English Language Arts	85	86	83	80*											
Written Composition	95	91	75	82*											
Maryland, Grade 9															
Writing			88	83	82	67	**69**	**54***	**51***						
Citizenship			75	76	71	73	66	**59**	**42***						
(Statewide data on the field tests in math and reading are not available)															
Montgomery County, Maryland, Grade 9															
Mathematics			82	83	84	85	86	83	79	**78**	**65***				
Blacks			61	63	65	64	67	65	57	**53**	**34***				
Hispanics			62	61	67	68	64	63	66	**61**	**42***				

	1991	90	89	88	87	86	85	84	83	82	81	80	79	78	77
Reading		96	97	98	97	97	97	98	97	96	92	92		90	89*
Blacks		93	93	95	94	94	96	95	93	90	83	79		72	66*
Hispanics		85	88	90	87	87	86	92	89	87	83	81		84	78*
Citizenship	85		84	81	83	81	75	62*							
Blacks	73		68	63	67	64	56	36*							
Hispanics	63		67	61	64	61	58	42*							

* Field tests or other "no fault" tests. The other tests, not starred, are required for high school graduation. Only first testings of each group of students are shown, not re-testings.

Source: State Departments of Education, and Montgomery County Public Schools [8]

NAEP tests are penalty-free, like the "no fault" tests starred above. A junior high school teacher told me of watching students on standardized tests fill in box 1 on question 1, box 2 on question 2, etc. in neat diagonals down the page, or drop the pencil randomly on the answer sheet. When she asked them why they didn't at least try to answer the questions, they asked "Why bother?" and she had no very good reason to offer. A junior and senior high school principal says the schools don't know how to get most students to take seriously any test for which there is no penalty. Both the teacher and the principal said 8th grade is especially not a good year to get students' cooperation, and February not a good month, so the trial state math test is doubly damned.

The introductory script read to students in the math tests [9] does not offer any strong reason why students should try hard. It says, "the results will help government leaders, school administrators, and teachers" (not the favorite people of all students) and "will have an impact on schools and students," (vague?) so "we hope that you will do the best that you can." The script goes on to teach students how to use a scientific calculator (where the order of key strokes may be backwards from what students are used to) by 4 examples: 4 x 7.3 - 2, (80 - 14) x 6, 29, and pi. This is not what most educators would call a thorough lesson. Then there are some sample problems, including algebra, which is likely to frustrate students who have not studied algebra. Then there are personal background questions [10] which end with, "Does either your mother or your stepmother live at home with you? ... Does your mother or stepmother work at a job for pay?" and similar questions about "either your father or your stepfather." I understand researchers'

interest in these questions, but the topics of divorce and stepparents are very touchy for many 8th graders and may leave students tense during the test itself. Then the third to last math background question is whether they agree or disagree that "mathematics is more for boys than for girls." Girls faced with this question may legitimately get angry at the presumption of posing such a question.

Overall, motivation may not be high when students start the test. There is a special problem for 12th graders: 42% are not taking any math [11], so many of them have little interest.

Curriculum Alignment

There is another factor present in the state graduation test results, with relevance for NAEP. As these tests became required, and teachers realized that the tests would actually be enforced as graduation requirements, teachers taught more carefully the material that would be tested. This accommodation shows up particularly in Maryland data, where the kinds of writing and legal knowledge that are tested were not necessarily taught throughout the state before the tests were required [12]. A high stakes test gives the test designers great power to control the curriculum [13].

Any national test that became a graduation requirement or job requirement would have a similar effect standardising the curricula, as SAT and ACT now do for college prep courses. The country will have to think whether it wants this standardization. For example the 1990 NAEP math test in 12th grade gives 45% of its weight to geometry and algebra. These go beyond simple applications like area equals length times width, to include secants of circles, supplementary angles, conic sections, imaginary numbers, and the quadratic formula [14]. For clarity of reporting, the objectives should be printed in the final report, so the press and public know what the students were expected to know. These math objectives follow recommendations of the National Council of Teachers of Mathematics, but are at odds with some minority views [15]. They are also at odds with skills listed in the last NAEP test on career development [16]. The goals of each test are set primarily by a group of college and public school teachers in the field, who have no special expertise on what the general population's needs will be in the 20th or 21st century.

On the writing test, NAEP collects 7½-45 minute writing samples [17]. On the other hand Simmons [18] found that poor students needed to put 16 days into their writing (though not full time!), compared to 13.3 days for the best students and 11.9 days for average students. With this amount of work, the poor students rose to about the middle of the class, instead of being much lower, as they appear on timed tests. If the NAEP writing test became a high stakes test, teachers and students would have to practice 7½-45 minute writing samples (with no time for reflection or re-writing). This writing drill would be at the expense of longer work, and also at the expense of speaking and listening skills, which already get little teaching, and yet are more central to "world class" workers than fast writing is [19].

Calculating Student Proficiency

The appendix to this paper explains how NAEP reviews the pattern of answers to the test questions. It explains that each student has an unknown proficiency on each topic in a test. Therefore NAEP does not estimate one score, but 5 likely scores for each student on each topic tested. We can consider these 5 proficiencies as 5 shadow students, each with a different score. The shadow students are intended to be a representative sample of all students.

In the 1990 math test, there were 5 topic areas as well as the 5 shadow students. Thus each of the 5 shadow students had 5 topical scores. These were averaged to create an average math score for each shadow student. NAEP reports show what fraction of shadow students are above or below various cut-offs, based on these average scores, or based on the 5 sub-scores. The percentages are of no great interest, since the scales are set to ensure that about 50% of students are above 250, 17% are above 300, and 2½% are above 350. The issue is what knowledge the students at each of these levels have, that others do not.

Describing Problems

NAEP publishes a curriculum simultaneously with administering the test. However they do not rank this curriculum from easy to hard. They wait until the test results are in, and then see what types of questions the students at various levels tended to get right and wrong [20]. NAEP then has groups of educators in the field try to describe the questions in terms of general kinds of knowledge (e.g. simple algebra). This procedure is hard, since there are overlapping concepts, questions worded in difficult English and questions surrounded by other harder questions. Then NAEP shows findings about how many students have each kind of knowledge. NAEP does not interview students, so it never knows why they get wrong the problems they do [21].

To show this process more specifically, we return to the 1990 math test. As mentioned above, each shadow student had 5 scores in different topics, which were averaged to get an overall math score. Then NAEP looked at shadow students who had average scores between 187.5 and 212.5, and found what percent of them got each problem right. Problems that at least 65% of these students got right (and that at least 100 students attempted or skipped) were considered fairly easy and

were used to give examples of what most students can do who scored at 200 or above ("anchor" problems). A group primarily of math professors and teachers looked at these problems and described them as **"simple additive reasoning and problem solving with whole numbers"** [22]. They also wrote a longer description which mentioned that these students can multiply and divide with a calculator [23]. For 8th graders they released 5 of these level 200 problems. The 5 problems included knowing a common factor of 10 and 15 (division without a calculator) and solving (150 ÷ 3) + (6 x 2) (multiplication and division which the authors thought would be done without a calculator) [24], so we have to be concerned by the short title which implies these students know no multiplication or division. The longer descriptions are not included in the executive summary and were not used in news reports. They were not included in the Education Department's own article on the results [25] and are only available in the $28 full report, the technical report and the state reports.

NAEP also looked at shadow students who had average scores between 237.5 and 262.5, and looked for problems that at least 65% got right, but which 30 percentage points fewer of the shadow students at 200 ± 12.5 got right. The same group primarily of math professors and teachers described these problems as **"simple multiplicative reasoning and two-step problem solving"** [26]. Their longer description mentions "factor" and "evaluation of simple expressions" in algebra [27]. Similar steps resulted in problems typical of levels 300 and 350, and descriptions of these levels. The short title of level 300 includes the words "simple algebra." They mean work more advanced than is done at level 250, but the brief titles wrongly imply that no algebra is done at level 250, just as they imply no multiplication is done at level 200.

The present anchor items describe an average of 5 math scores. Each level may include students good in statistics but bad in algebra or vice versa. It would be more meaningful to describe anchor items for each subscale separately.

The task of describing common patterns of what students can do is very hard. Often similar problems have very different success rates, and it is hard to see a reason. Neither NAEP nor the news reports highlight how ambiguous it is to try to say what a group of students can do, based on a few test questions. Right answers may often depend on the context of questions [28]. Several of

the harder 8th grade anchor problems come from a single block of questions that students found hard (41% of problems in this block were answered right on average) [29]. The block started with a question on converting 150 minutes to hours, then had an algebra problem and a solid geometry problem. It had several other hard algebra and geometry problems, which may have frustrated students. Lord pointed out that the presence of hard problems hurts performance even on easy problems, since the hard problems take students' time away from the easy problems [30].

There are other examples of the problem of describing in words what students can do. In the 1988 writing test, students were asked to write a persuasive letter. The assignment and the criteria were described quite differently in two reports on the same test [31]:

Assignment:

1/90 report: "adopt a point of view about whether or not funding for the space program should be reduced, and to write a letter to their **senator, explaining their position.**"

6/90 report: "take a stand on whether or not funding for the space program should be cut and write a **persuasive letter that would convince a legislator** of this stand"

Criteria for minimal:

1/90 report: take a point of view, **not present reasons,** no convincing evidence to sway senator's vote

6/90 report: take a stand, briefly support it with **one or two relevant reasons**

I have been told that the same test question and scoring criteria were being described in these two reports, one on 11th graders, the other on 12th graders [32]. The 6/90 version changes the tone of writing expected and hides a flaw in the test for Washington DC students, who had no senator (the "Dear Senator" seems to have been pre-printed on both answer sheets). The changing definition of "minimal" makes the results impossible to interpret. The definition of minimal is key, since half the 11th graders are at this level. Actually either definition should probably be called better than minimal, since lobbying groups recommend a simple brief statement of one's stand [33]. There is an air of unreality about the assignment anyway: 7½ minutes to convince a senator who has been the target of large professional lobbying campaigns? Nor did either report mention that the time available was 7½ minutes.

The IAEP report on math and science gives even less information to judge what the different score results mean, with only one problem at each scale level in math and science [34].

The NAEP staff undoubtedly try to present clear explanations of what is known at each ability level. The task may be impossible, especially with students learning different aspects of writing, math, listening, lobbying, etc. in different schools. The report needs to mention these difficulties.

With the 1991 math report, NAEP has made a large improvement in presenting information on math achievement. Up through 1988, reports showed how many students scored at and above various scale values, but did not mention that many other students also answered right each of the problems presented as typical of the scale value (since some students at lower levels also get each problem right) [35]. Now NAEP shows what percent of students get each problem right and the press reports it.

Aside from the difficult descriptions of the levels, NAEP now presents the percent of students scoring at or above each level in meaningful ways. The report talks about students "demonstrating the ability" or "consistent success" or "solid grasp" [36]. These terms are fairly meaningful. Students at each level score about 70% on the problems typical of that level. The problems are independent, so students do not have a 70% chance of getting them all right, but on average they will get 70% of these problems right. Typically about 30% of the students one standard deviation lower get each of these problems right. So those lower students show a weak grasp, or inconsistent success. By contrast the report on the 1986 math test implied that a level was all or nothing: students knew the skills at a level or they did not, which led to the mistaken belief that the percent who could do a problem equalled the percent who were at that level [37]. One change that would help would be to avoid saying what students <u>can</u> do, based on the test, and say simply what they did. As noted above, it is very possible they <u>can</u> do more.

HOW THE PRESS REPORTS NAEP

For this paper I reviewed 15 news accounts of the 1990 NAEP math test, and a few accounts of other tests [38]. The press reports are mostly very similar. The headlines usually say students are failing (9 out of 15). The text repeats some main numbers from the NAEP report (or from the SAT, ACT or norm-referenced test) and some quotes from education professionals who have ideas about what should be done. The ideas may change, "choice" in 1990, American Achievement tests and new math curricula in 1991, but the pattern of the stories is fairly constant.

The result is not necessarily a consistent push for a needed reform, but a general belief that students, parents, teachers, textbooks and bureaucrats are no good, creating poor morale, especially among teachers, without the detailed information that would let someone know what improvements to consider.

The newspapers generally do very direct reporting of NAEP results and the accompanying political statements. They report average scores, compare various groups, and quote the interpretive statements provided. "Where will the world's innovative discoveries, new solutions and creative products come from in the future? Does it matter?" was quoted from the IAEP report on math and science in the Boston Globe [39]. "How many times must this nation be reminded of its educational deficits?" was quoted from Secretary Cavazos in an AP story in the New York Times [40] on the same IAEP report. "Students are generally ill-equipped to cope confidently with the mathematical demands of today's society, such as the graphs that permeate the media and the regulations and procedures that underlie credit cards, discounts, taxation, insurance and benefit plans" quoted the Richmond Times-Dispatch from the 1991 math report [41].

The papers generally said most students were not ready for college (11 of 15) or technical jobs (8), and that 8th graders largely can't do fractions, decimals and percents (10 of 15).

None of these papers covered any of the following on the 1990 math test:

Comments from alternative test proponents, such as the supporters of portfolios and performance assessments

Comments from 8th grade teachers or students

Caveats such as lack of student motivation, average response levels of 80% (down to 62%

in Oklahoma), varying percentages of students omitted because they were in private schools, small numbers of problems, unfamiliar scientific calculators, etc.)

The issue of whether algebra and geometry should have 40% weight in the 8th grade, though these are often not taught by then

The trial nature of the state testing, with its meaningfulness still in doubt

Reliability of NAEP descriptions of scores and student failure

The reporters thought this was a fairly straightforward story, repeating widely known problems. They trust NAEP to have large sample sizes (mentioned in 7 of 15 stories), well spread around the country [42]. They have little knowledge of psychometric difficulties in interpreting what students know.

Only one story that I saw had a substantially different interpretation from the NAEP report itself: the Wall Street Journal said, "States with traditional classroom approaches ranked highest in the study." [43]. The two reporters who wrote this article were able to find this information in the NAEP data and decided for themselves that it was a significant finding.

The reporters are generally capable of covering more of the issues on testing and the math curriculum, even on small newspapers. However they seem to do thorough coverage primarily in feature stories, which may develop over time. Newsweek did straight reporting of this test. Time did not, but may work it into some more general story in the future [44]. The reporters occasionally cover stories on opposing viewpoints, such as a story in the Bismarck Tribune that extracurricular activities are predictive of later success in life, while school and college grades and ACT scores are not [45] and a story in the Atlanta Journal that US adults know more science than Japanese adults [46]. However the authors who release such reports generally lack the publicity resources of NAEP and get much less coverage.

Time for Reporters to Understand the Issues

In talking to reporters about the coverage of the 1991 math study, they complained that the materials were voluminous, and they did not have time to digest them [47]. The press received an advisory several days ahead that the report was coming out. It did not say so, but the report

itself was available noon the day before the press conference, under an embargo. For a 500 page report, that gave the press little time to understand it [48].

The reporter for the smallest paper I spoke to, the Bismarck Tribune, said she needed at least a week and preferably two, under embargo, in order to understand the report, get comments from teachers and make the story a local story. Even on the day of the press conference, officials in her state said they only had two copies of the report and refused to give her one [49]. Larger papers did get the report, but also wanted up to a week, also to understand the report and explain it better. NAEP worries about a longer period of embargo, saying the results were so sensitive they had to be held very tightly [50]. The press did not seem to consider the results so sensitive, since most states differed little anyway, and the overall results matched the conventional view that students are doing badly. They did not worry that someone might break the embargo. On this report, the Boston Globe did break the story a day early. Papers worried most about competition with TV news in their own markets, and they already lose that race with evening TV news, when NCES releases the information at a morning press conference. I think that a longer period under embargo would result in better coverage, and the occasional leaks would cause little harm.

Statistical Presentation

Several reporters mentioned that NAEP wanted them to use the "pantyhose" chart from page 16 [51], to show in a statistically sound way which states outranked which others. It lists all 40 jurisdictions tested, down the side, and lists them again across the top. For each jurisdiction one color shows which other areas are statistically the same. Another color highlights the states that scored better (or worse) to a statistically significant degree. The reporters thought such a chart was unreasonable for a newspaper, and wanted a simpler presentation. Some papers listed the states in alphabetical order, with scores and ranks. Some listed them in rank order. The Des Moines Register and New York Times listed the top and bottom states. Newsweek and the New York Times showed visually in bar graphs how much and how little the states differed. My impression is that the reporters and probably the public had little interest in the details of the ranking, aside

from top, bottom, middle. Selden (52) suggested a graphical relation of socioeconomic status of states to test scores. The papers might carry such a graph, but the reporters and probably readers would still think the bottom states ought to be improved and the top states probably also, as Mr. Selden accepted in his article. I also think in his article Mr. Selden thought there would be more statistically significant differences among states than there turned out to be.

The scale of proficiencies from 200 to 350 was not easy to understand. Newsweek was boldest, stressing which grade each score was equivalent to. They went beyond NAEP's careful statement that level 300 material is "introduced by the 7th grade" [53] to say "300 is roughly seventh-grade work." Some might disagree anyway on whether $2x + 3y + 4x$ is introduced in 7th grade, or the inequality sign in $2x > 11$ [54], though many of the other items are more clearly 7th grade work.

Cause and Effect

NAEP reports do not try to measure cause and effect, and newspapers generally preserve that line. The papers usually mentioned some correlates of the scores, especially TV time (14 of 15 papers), race (12), 2-parent families (12; neither the papers nor the executive summary mentioned that these included stepparents), parents with college education, sex and suburb/city comparisons (9 each), attendance and poverty (7 each), home reading materials and homework (5 each). On that list, schools have some control over homework, but otherwise the aspects that schools can control were mentioned rarely: ability groups, school budgets, computers and workbooks were only mentioned by 2 papers each. This pattern reflects the stresses in the NAEP report.

Nevertheless I would not encourage papers to give more play to correlations between scores and school actions, since the correlations may be spurious. One would first need to look at each effect while controlling for others (in a regression), and one would still have to deal with the ambiguity caused by lack of student motivation. For example perhaps ability groups result in lower test scores only because they reduce school loyalty and therefore reduce motivation on this kind of a voluntary test, while they may have no effect or a positive effect on actual learning. Or there may be other spurious connections between ability groups and test scores. There is certainly active research on the effectiveness of ability groups and other actions schools can take. NAEP is

probably not the best place to study that kind of specific issue. The same weaknesses apply to the demographic issues that do get wide play. As a first step, NAEP can report on multivariate analyses to see what contribution each of the variables makes to math proficiency (or at least to test scores) when one controls for the other variables. Presenting such information is certainly feasible for newspapers. They can use concepts like: x points are added to a score by daily use of calculators, y points are subtracted for each hour of daily TV watching, etc. This multivariate appproach, in combination with Selden's graph, might encourage people to see which states are doing better than their socio-economic status would suggest, so other states can copy what they are doing right.

Splash

NAEP reports editorialize more than many government press releases, in order to make a splash. The Labor Department says, "The nation's employment situation was little changed in June ... The unemployment rate was 7.0 percent, little different from the May level of 6.9 percent" [55]. The Department of Health and Human Services says, "mortality rates for ... hospitals [were] released today ... consumers should use the information in consultation with their physicians. Mortality rates ... do not necessarily represent the total performance of a hospital in caring for its patients" [56].

On the other hand NAEP reports have such phrases as, "a large percentage of students approaching high school graduation ... lack a sense of the national heritage" [57]. "The mathematical skills of our nation's children are generally insufficient to cope with either on-the-job demands for problem solving or college expectations for mathematical literacy" [58]. Yet half of high school graduates do go on to college, and seem to cope; and the 20-24 year old unemployment rate seems to have little connection with state by state test scores, so people seem to cope at work too [59].

NAEP says the US "is having difficulty maintaining its competitive edge in the global marketplace" [60], though our productivity is $24.29 per worker per hour, while Japan's is $12.76 [61], and anyway in a service economy, most workers are not in danger of their jobs moving abroad. NAEP also complains that only 800 students get doctorates in math each year, down from the baby boom years of the 70s [62]. The relationship of global competitiveness and doctorates to some of the math questions covered in the report seems tenuous. Perhaps NAEP believes its data are less significant than the unemployment rate or the hospital death rates, so they have to color their language [63].

IMPROVEMENTS IN NAEP REPORTS

The NAEP reports would be clearer and have clearer news coverage if they had a longer period of release under embargo and if they had a three page summary, with one page on each of the following:

Main findings

Source of the data

Limitations of the data

The first two topics are covered in the present NAEP reports, but limitations are not, so I will list some of the items I have in mind:

Most 8th grade students have not been taught some of the topics tested, such as algebra, geometry and probability (totalling about 40-45% of the total score at grade 8) [64]. NAEP does not seek to impose a national curriculum and therefore does not recommend that schools try to improve scores by teaching they topics they do not want to.

States also vary in students' educational backgrounds and family lives (such as the amount of quiet, stability, and encouragement the students have at home). Therefore some states have a harder time than others in teaching even the same material.

The results are biased downwards to some unknown extent, since the test is voluntary, so students have no incentive to do their best. Differences among scores may be caused by differences in students' willingness to devote energy to a voluntary test.

In NAEP and any test it is very hard to summarize in words what it is the students can do.

Response rates vary, with 80% responding nationally at 8th grade, or as low as 62% in Oklahoma [65] or 65% in 12th grade nationally [66]. Coverage rates are lower than response rates, considering the omission of private schools, non-English speaking students and special education students.

The test scores have not been proven to have a relationship to success in later life ("predictive validity").

A summary of limitations like these would give the press and the public some orientation to the

data. Similarly the Department of Health and Human Services in its press release on hospital deaths mentions caveats, in a way simple enough for reporters to cover (this example was suggested by Jane Norman of the Des Moines Register) [67].

The backup sections in the report would include more detail on each of these sections, and:

Detailed objectives, i.e. the content intended to be covered by the test

Proficiency on each topic, among students who have been taught that topic

Actual released tests, with accompanying scripts, percent of students choosing each answer or omitting the question, and a, b, c parameters (see appendix)

Regression coefficients or other multivariate information, showing the effect on performance of each background variable or cluster of variables, holding the other variables constant; this would largely take the place of the univariate statistics now in NAEP reports.

Non-participation rates, combining student and school non-participation rates, and also overall coverage, considering special education, language barriers and private schools

It would also enrich the reports if NAEP could study students' attitudes and thinking processes as they take the tests, by observation and by interviews. This is a field where cognitive psychologists, child psychologists and anthropologists could be helpful [68].

The assignment of grade equivalents to NAEP scores seems very unwise, since curricula can and should vary: algebra may be taught in one school in 7th grade and may never be required in another school at any grade. To assign any grade equivalent is to assume a certain curriculum, which is not NAEP's role.

Overall, considering the press coverage of NAEP, it is hard to see that the taxpayers are receiving information commensurate with the cost of the NAEP tests, and especially the state assessments:

The tests do not cover the major issues generally agreed to be needed in work and life: teamwork, work attitudes, speaking and listening skills, etc.

The content of the tests is not and perhaps cannot be summarized accurately

The students lack motivation on NAEP tests, and state differences and changes over time are within a range that could be explained by differences in motivation

A test with enough sticks or carrots to create motivation would move control of the curriculum to the test-writers

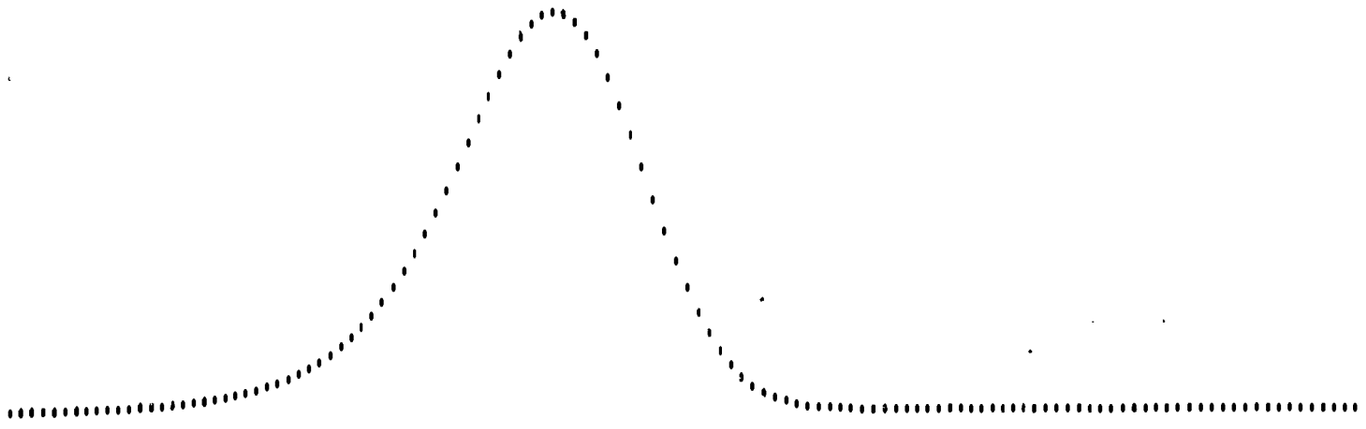
Most countries do not even try such general tests in their high school examination systems. They tell teachers and students years in advance which topics will be tested, give them strong incentives, present students questions, usually with a fair amount of choice, and note whether the students display a serious understanding of the chosen problems, without trying to generalize to broad topics [69].

The US does not need tests to make schools accountable. As with doctors, judges, artists or mechanics, the difference between good and bad is not a score on a test, but is a complex matter, often different in the eyes of different beholders. Qualitative comparisons of schools, by various groups, such as newspapers, parents, students and businesses, would be richer and could focus on important differences of atmosphere, teaching ability and broad learning, more than test scores do.

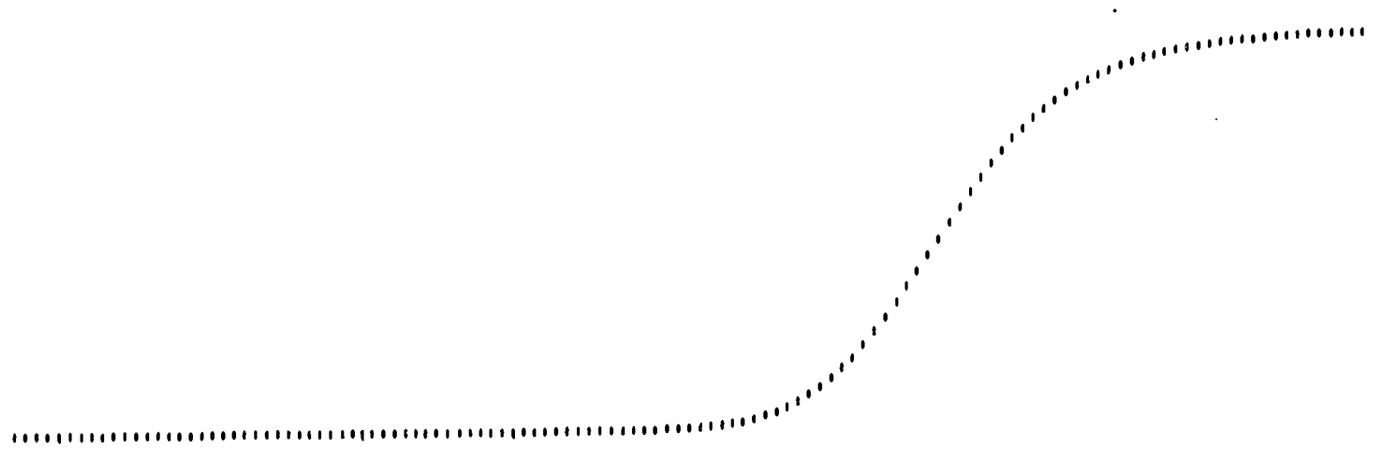
APPENDIX

Calculating Students' Scores

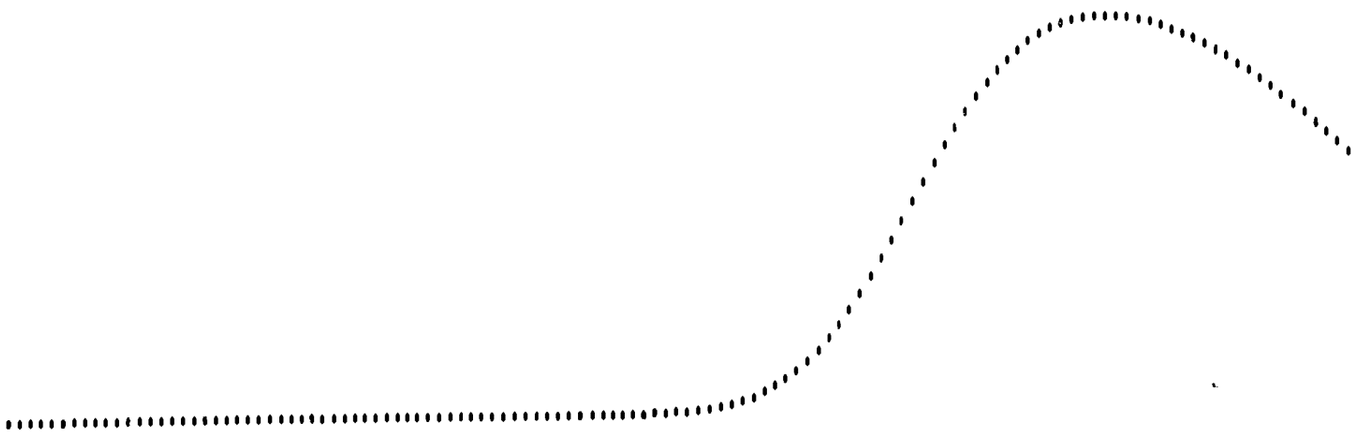
One of the purposes of this paper is to explain how NAEP estimates the likely "proficiency" of students in NAEP tests. NAEP gives each student several questions on a particular topic. As an example we can look at the 7th booklet of the 1990 math test. It has 8 problems on data analysis, statistics and probability [70]. As mentioned earlier, NAEP looks at which problems students get right and wrong, to estimate their "proficiency" or "display." For example within these 8 problems, students who get the easiest 4 right and the hardest 4 wrong, are likely to be distributed in their proficiency according to the following curve [71]:



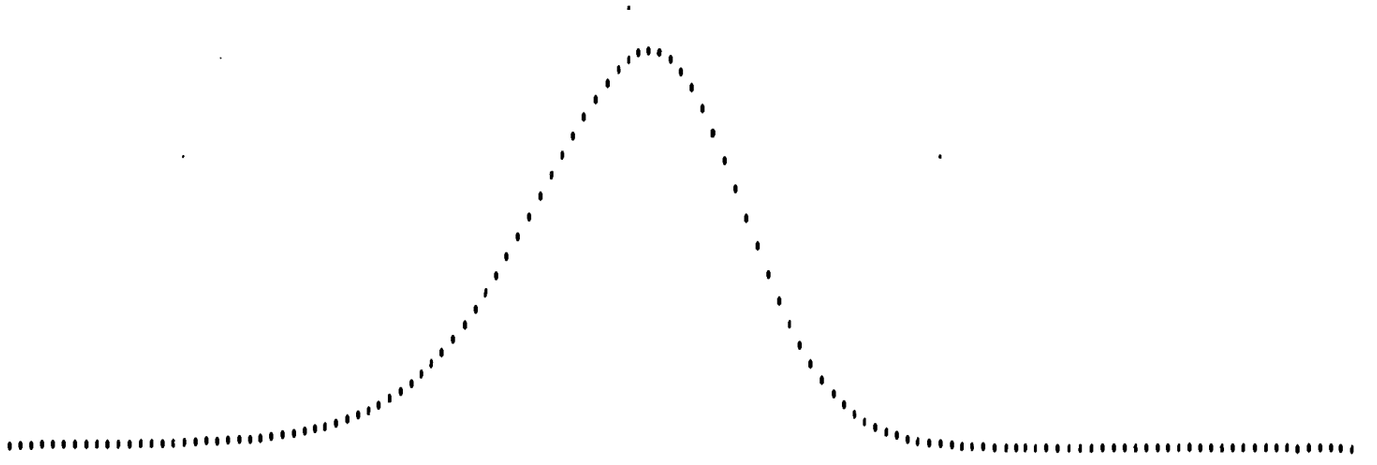
At the far left or right, a few low or high students may accidentally get the easiest 4 right and the other 4 wrong, but most students who have these 4 right and 4 wrong answers are likely to be middle ability students. Students who get all 8 answers right are likely to be distributed according to the following curve [72]:



However NAEP finds that curve hard to work with, so they use the following curve instead [73]:



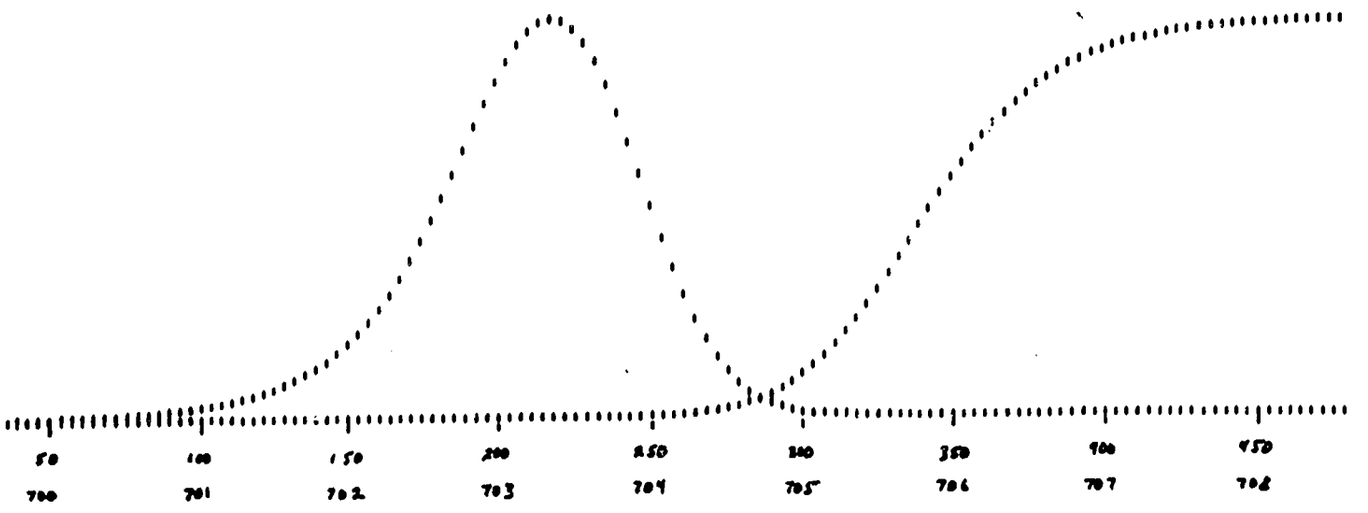
This lowers the scores of top students to be closer to middle students. They do a similar change at the bottom. Other students might get 4 problems right and 4 wrong, but out of order, say they get wrong the 2 easiest and the 2 hardest. Such students are likely to be distributed in their ability according to the following curve, very similar to the first curve:



I haven't yet labelled the scale of student abilities from left to right. I haven't explained how these likely distributions of students are figured out. I haven't explained how the distribution of the total population is figured out from these distributions for individuual patterns of scores.

Scaling

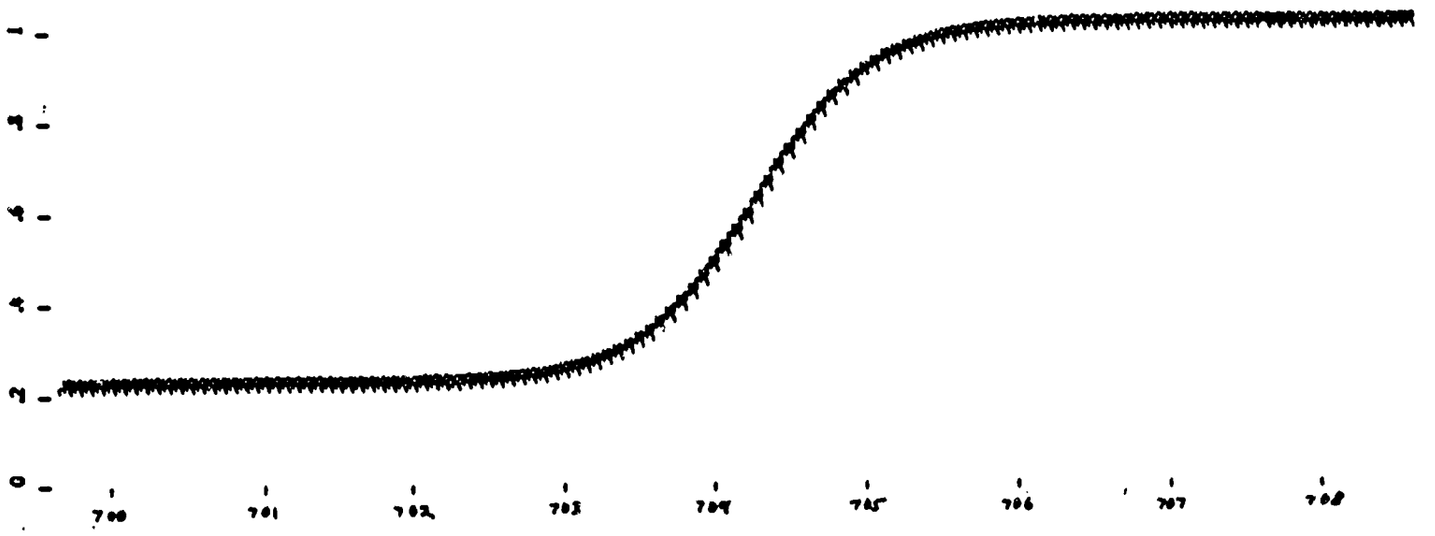
I haven't labelled the scale, since it is arbitrary, in the same way that Fahrenheit and Centigrade temperature are. The zero can be anywhere, and the steps can be any size [74]. NAEP takes the average of all students in 4th, 8th and 12th grades, and calls it 250.5. They take the standard deviation of these students' proficiencies and call it 50 [75]. If the scores followed a normal bell-shaped curve, 17% of all students would be below 200, 17% above 300 and 2½% above 350. In fact only about 10% are below 200, 22% are above 300 and about 2% are above 350 [76] taking all grades together, so the distribution is slightly skewed. As an alternative scale, one could label the scores with the average at 704 (Independence Day), and a standard deviation of 1. One would still have about 2½% of students above 706. With these labels the first two curves above would be:



The scale from 700 to 708 creates a subtle impression that there is not much difference in knowledge. The scale from 50 to 400 implies that people at 400 know twice as much as the people at 200. On a vocabulary test it may be meaningful to know twice as many words (though there are rapidly diminishing returns, since 8,000 words account for 90% of written English, and knowing another 8,000 words only accounts for another 5%; [77]). On most tests there is no obviously meaningful scale, and the reader must guard against thinking that 400 is twice as good as 200. NAEP tests are designed to distinguish students from each other, not to measure what they all know. NAEP reports themselves never make the mistake of interpreting the scale in terms of percentage differences, but they do not always say how arbitrary the scale is, and newspapers do say things like "Georgia ranked ... 30 percent from the bottom" [78].

Probability of a Right Answer to a Specific Question

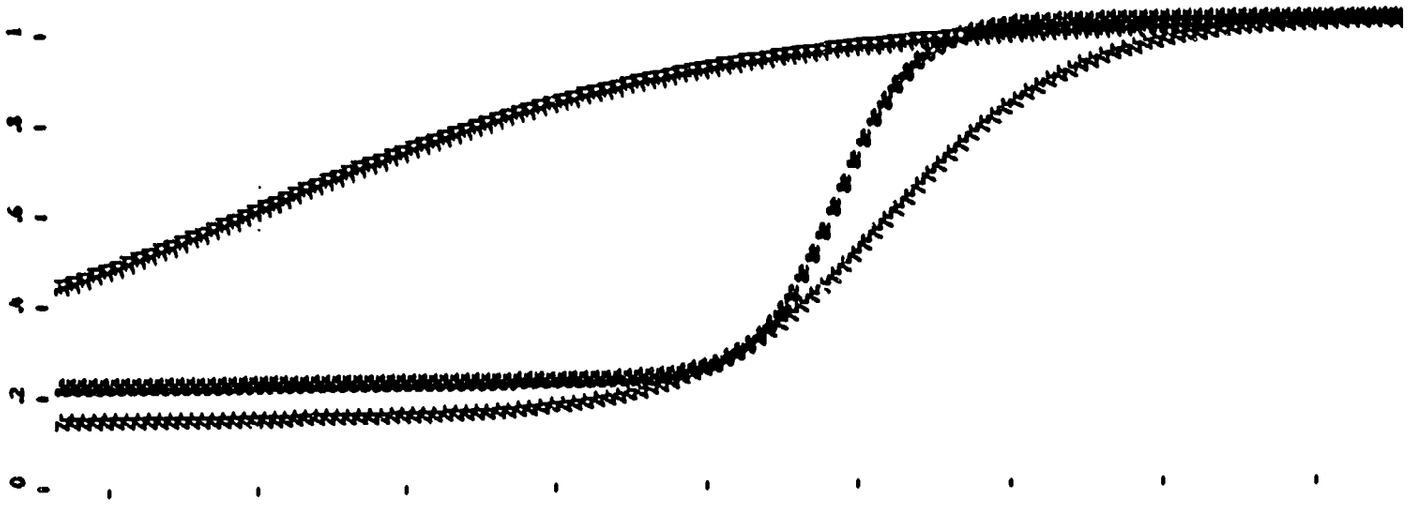
We need to analyze each question before analyzing a whole pattern of answers to questions. NAEP fits a mathematical curve to each question, showing the probability of a correct answer from students at different proficiencies. Here is the curve for a question on calculating the average age of 5 children: 13, 8, 6, 4, 4, with multiple choices: 4, 6, 7, 8, 9, 13, don't know (no calculator available) [79]:



Low students have a 22% chance of getting the right answer, so they're doing better than random guessing. High students have very good odds of getting the right answer. The problem is pretty good at distinguishing between low and high students, but not between low and very low or high and very high students, since the chance of a right answer is not very different once you get below 703 or above 706. For making small distinctions between similar students, the problem is best between 704 and 705, since the chance of a right answer improves fairly fast in that range. In fact the problem has a steeper slope than most, perhaps because it is near the end of its test, so it is measuring both speed and knowledge. (It also measures agreement that medians are not averages, which may trouble some. NAEP accepts 7 but not 6 as an answer. In the 1990 objectives book, authors of a similar question thought they needed to say specifically arithmetic mean when they asked for an average [80], so it is not clear why the authors here thought the term average by itself was unambiguous.) The equation of the curve is [81]:

$$p = c + (1 - c) / (1 + e^{-1.7a(\Theta - b)})$$

where p is the chance of a right answer, Θ is a student's proficiency, e is the mathematical constant 2.7183 and a, b, c are calculated by NAEP to fit the curve to real data as closely as possible. For example on this question NAEP calculated c=.214 (the guessing level, a lower asymptote), b=.104 (the difficulty), and a=1.368 (the steepness). For open ended questions c=0. On the 1990 test of data analysis, statistics and probability, the difficulties range from -3.623 (easiest, the bar graph on p. 63 of the full report) to 1.183 (not released, but it involved median). The steepnesses ranged from .333 (gentlest slope, the graph on p. 63) to 1.983 (not released, but it involved interpreting a circle graph) [82]. To give a sense of the range of curve shapes, we show these three problems here:



	a	b	c	
Problem	Steepness	Difficulty	Guessing	
3d	.333	-3.623	.175	easiest problem and gentlest slope
8e	1.983	.788	.216	steepest slope
7r	.860	1.183	.140	hardest problem

The problems are identified by their block (from 3 to 9; blocks 1 and 2 were background questions) and by the question order within each block (from a to v, representing questions 1 to 23).

From the actual test results, NAEP calculates Θ, a, b and c, and there is room for error. The testing literature has articles critiquing various ways of calculating these figures and simulating the amount of error resulting. The following table from Mislevy illustrates the

problems, using simulated data, where it is possible to know what the true values are, unlike real tests, where the true values are never known [83].

	a		b		c	
Question	True	Est.	True	Est.	True	Est.
1	1.1	1.3	-.4	-.3	.11	.17
2	.5	.4	.2	.6	.19	.24
3	.9	1.1	-1.3	-1.0	.26	.27
4	1.4	1.4	-1.0	-1.0	.17	.19
5	1.5	2.4	-.3	-.2	.13	.14
6	2.5	3.4	-1.1	-1.1	.18	.18

Source: Mislevy, 1986

Besides errors in the parameters, curves may have different shapes from the equation assumed, with other bends and twists. There may be other important variables. NAEP recognizes that different curves may be appropriate for different states, but they derive one set of curves in order to "maintain an equal measure for establishing comparisons among participating jurisdictions." They recognize this may mean the measure fits the curriculum and answer patterns of some states more than others [84].

Probability of a Pattern of Answers

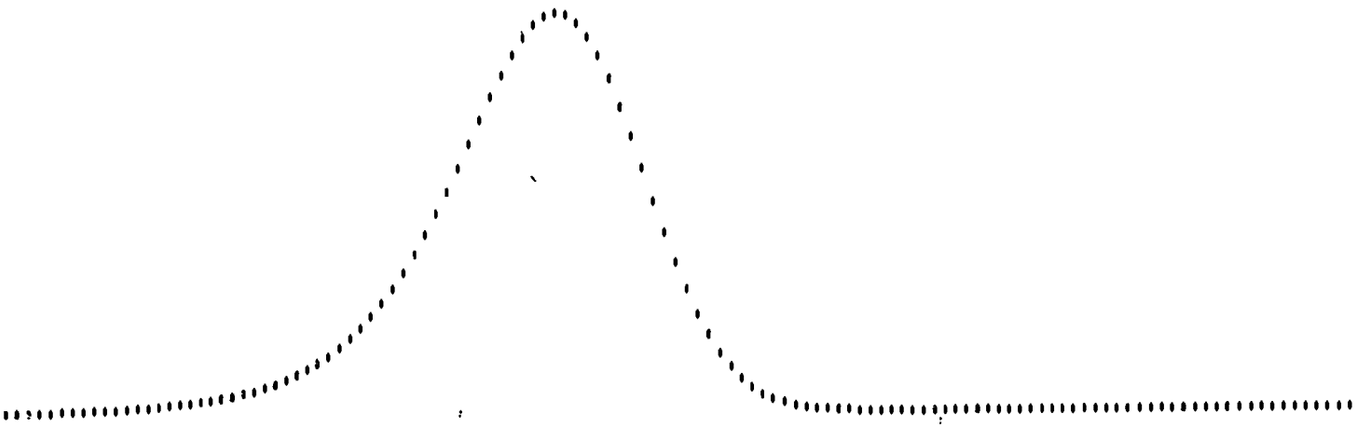
Once NAEP has an equation for each problem, the probability p of getting it right can be calculated for each Θ. The probability of getting the problem wrong is 1 - p. For each Θ, problems are seen as independent [85], and each can have its p calculated, as p_1, p_2, etc. In order to calculate the chances of getting two problems right we can multiply the two probabilities (just as the chance of 2 heads is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$). The chance of getting the first four problems right, and the next four wrong is:

$$z = p_1 p_2 p_3 p_4 (1 - p_5)(1 - p_6)(1 - p_7)(1 - p_8)$$

Remember each of these p_i depends on its values of a, b, c and Θ. A, b and c are fixed for each p_i.

We can choose values of Θ from low proficiency to high, calculate each p, then calculate z,

then graph the curve of z:



The curve is low on the left, since low levels of proficiency mean the probability of getting any item right is fairly small, so the product z is small. For high proficiency, the probability of getting any item wrong is small, so again the product z is small. In the middle, the probabilities of right and wrong answers are not so small, and the product z rises to its maximum. This curve is treated as the likely distribution of proficiencies for students who had this pattern of 4 right and 4 wrong answers on the test.

Combining Different Patterns into a Distribution for the Whole Population

NAEP does not simply add these distributions for all students. They recreate the total population by using various equations.

NAEP finds the mean of each distribution. Then NAEP tries to find one equation (a "regression") that calculates as many as possible of these means (for different students) as closely as possible. The equation takes into account background information on the student and the student's school (race, sex, parents' education, teaching practices, etc.) [86]. On average being black means fewer right answers and a lower distribution of proficiency, so does low parental education. So may certain teaching practices (though NAEP does not report their findings on the effects of teaching styles).

But of course not all students are at the mean proficiency of their group as calculated by the

equation, or even at the mean of their own personal curve. Students are scattered all over the curve. They are simply considered more likely to be in the larger parts of the curve. NAEP picks 5 proficiency values for each student, randomly ("plausible values" or "imputations") [87]. These are spread somewhat from the mean of their group, but not spread as much as the personal curves go. This is called a "posterior distribution," resulting from the "prior" assumption that students are more likely to be like other members of their groups than spread all over their own distributions.

There is also a "prior" assumption that all groups of students have the same dispersion (standard deviation). Mislevy et al. [88] state that these techniques preserve the mean, standard deviation, and shape of the distribution of proficiencies actually in the population.

How Reliable Are Short Tests?

The reader may have had a qualm when we mentioned that NAEP was analyzing a test with only 6 or 8 items. The qualm may have become anxiety when we pointed out that each question can distinguish detailed levels of proficiency in only a small stretch of the distribution. So at some levels, estimates of proficiency may be affected seriously by a single question.

There is a formula to measure the amount of information a test provides at each level of the proficiency distribution [89]. This is the amount of information available to distinguish one proficiency from another. The formula is:

$$I_1 + \dots + I_n$$

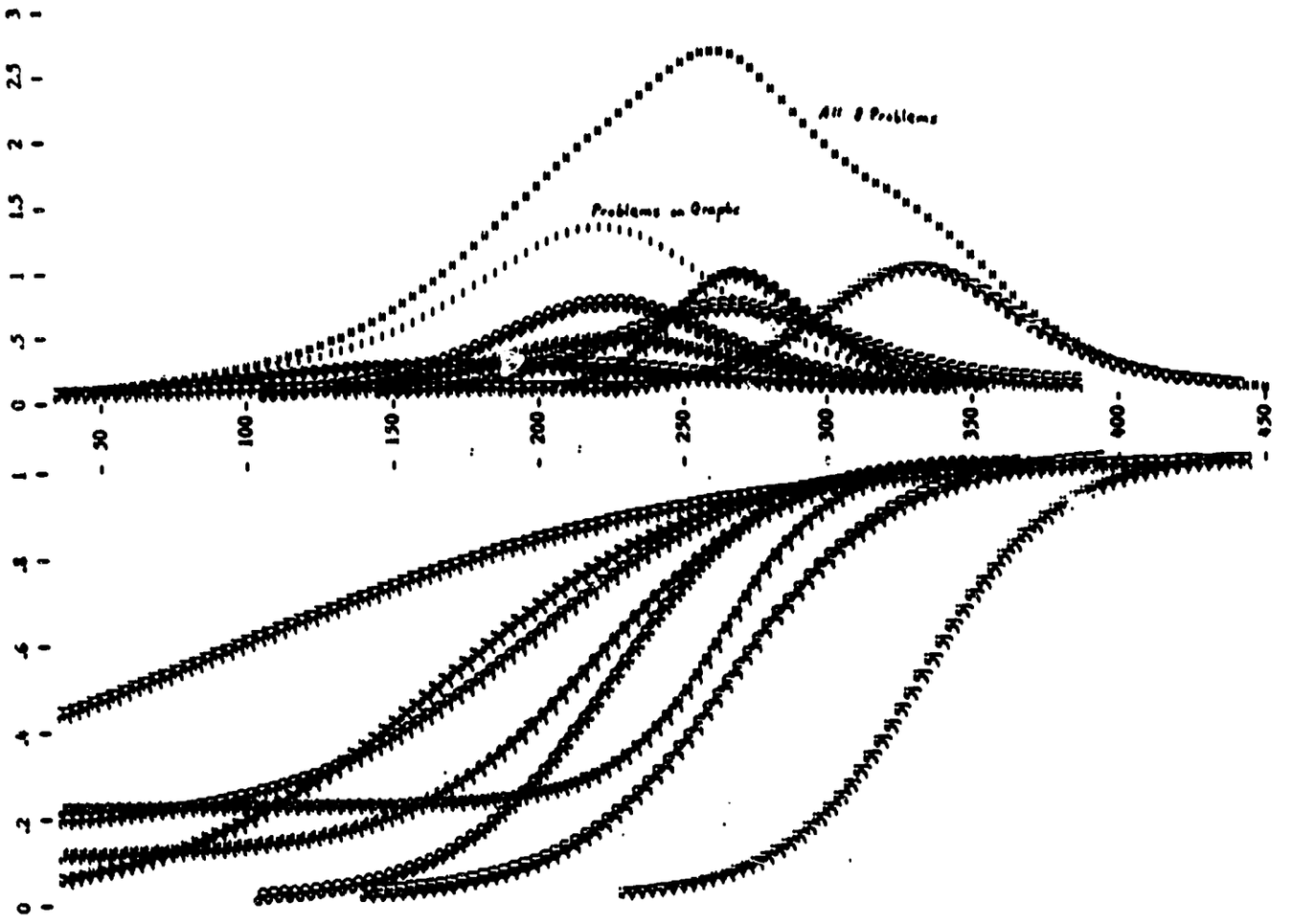
This formula adds up the information for all questions, where the I for each question is:

$$I = 2.89a^2(1 - c) / ((c + 1/k)(1 + k)^2)$$

where $k = e^{-1.7a(\Theta - b)}$. For the example of test 7 in the 1990 math test, with 8 problems, the top of the following graph shows the amount of information available from each question, the total information, and the information from four questions on graphs. The test includes 4 questions on graphs, 2 on probability, and 1 each on averages and sample bias [90].

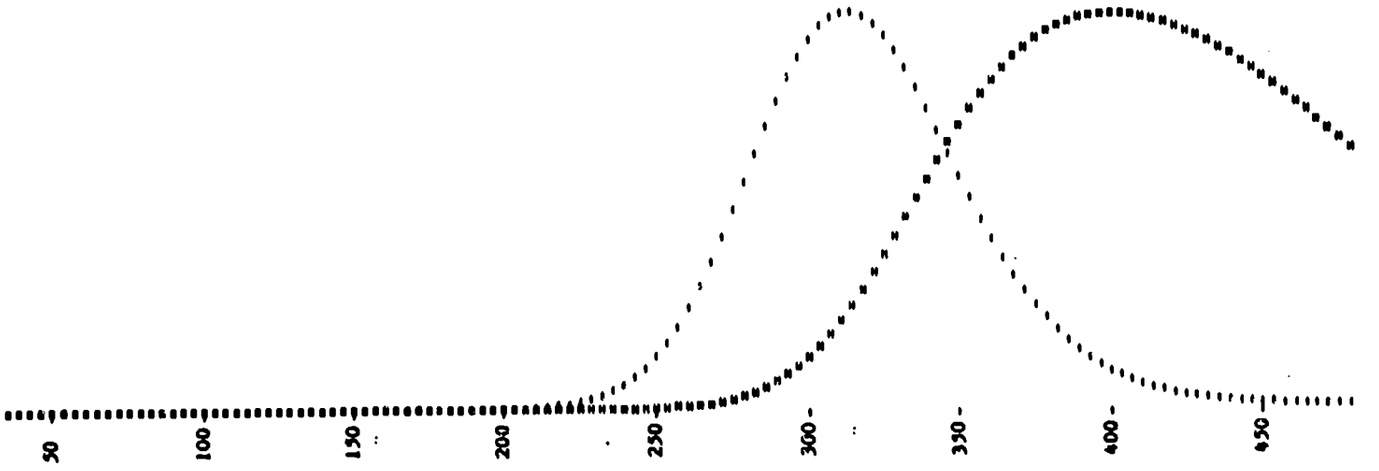
On the scale for amount of information [91], one means approximately the amount of information from one good question. The total information in the middle of the proficiency distribution is equivalent to about 2-3 good questions at each point. However at 350 there is effectively only one

question, which is the question on average ages shown earlier. The bottom of the graph shows the probability of a correct answer on each question separately [92].



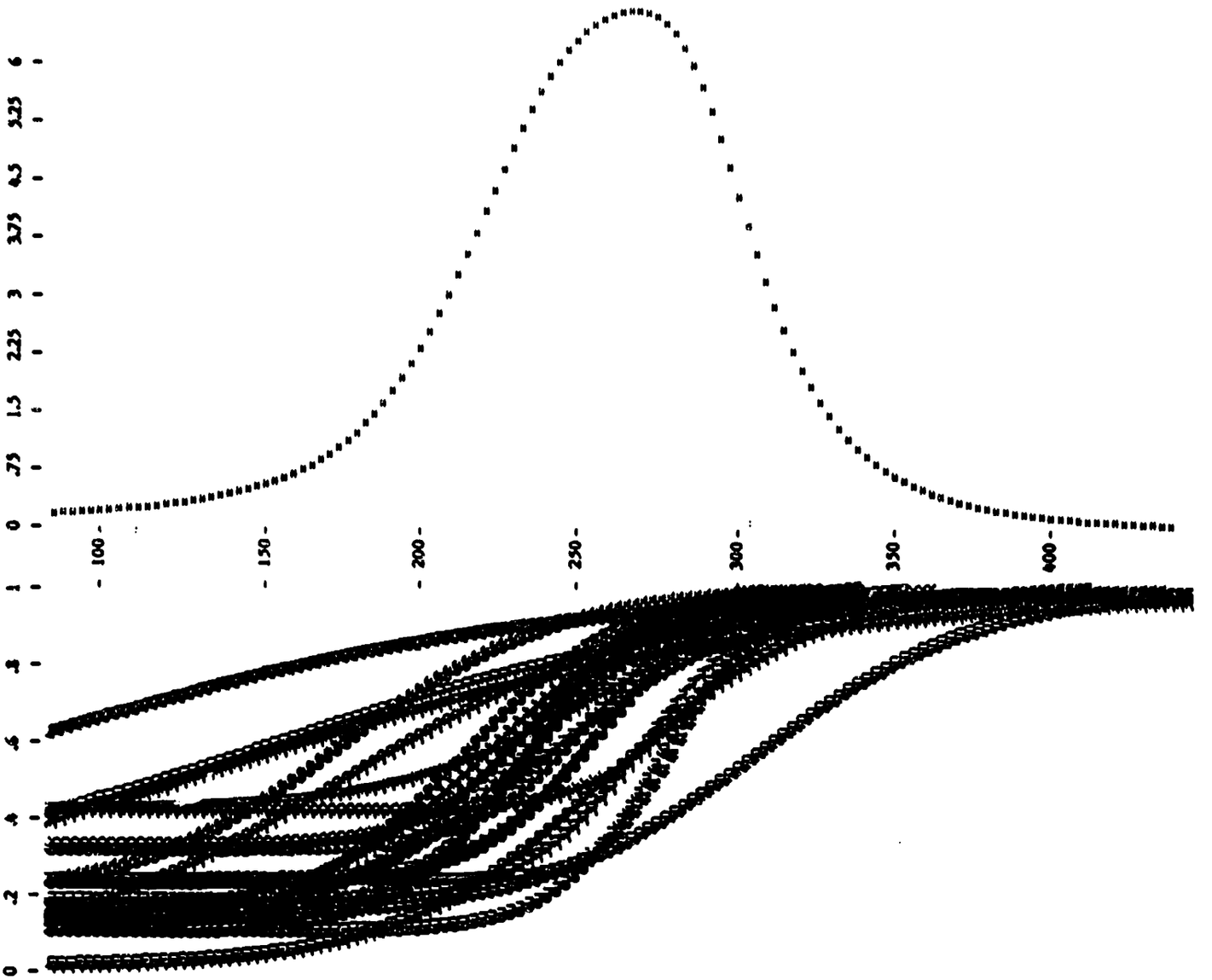
	a	b	c	
9p:	.981	-.777	0	draw + label circle graph, p. 67
3h:	.668	-1.437	.175	odds of 1 marble from a bag
3d:	.333	-3.623	.175	read bar graph, p. 63
3e:	.829	-.881	.104	compare on bar graph, p. 473
3s:	1.368	.104	.214	average age of 5 children, p. 477
5f:	.576	-2.059	0	complete a bar graph } not released
5j:	1.14	1.63	0	list sample space
5m:	.944	.157	0	explain sampling bias

The next graph shows the effect on a student's likely distribution if she misses that one question. The mean of the distribution drops from 400 to 320.



Student proficiency estimates depend seriously on a relatively few questions at the high end of the difficulty range, even when we consider all test booklets together, not just the questions posed to one student. If there are too few hard questions (or easy ones) to be a fully representative sample of the domains of mathematics that they should represent, the test is weakened. For example the 1990 subscore on data analysis, probability and statistics has only 19 questions in all, and only 5 with b parameters at level 300 and above [93]. At best one can see these 19 problems as a well-stratified systematic sample of a certain domain of knowledge. This is a fairly small sample size, from a large domain. It is possible that NAEP is trying too much when it tries to measure a wide range in 5 abilities in 45 minutes.

In the 1990 math test, the largest number of math questions were on "numbers and o\ ıtions" with a total of 46 questions [94], which is also a limited sample. As a further example of test information, booklet 4 had 23 questions on numbers and operations [95]. The following graph shows the total information available and the probability of a correct answer on each question. Again, little information is available around level 350.


6
5.25
4.5
3.75
3
2.25
1.5
.75
0
1
.8
.6
.4
.2
0
100
150
200
250
300
350
400

FOOTNOTES

1. Stephen L. Koffler et al., The Technical Report of NAEP's 1990 Trial State Assessment Program, (Washington, DC: U.S. Department of Education, National Center for Education Statistics, April 1991).

2. Geoffrey N. Masters, "Item Discrimination: When More Is Worse," Journal of Educational Measurement, 25:1, Spring 1988, pp. 15-29.

3. Ronald K. Hambleton, "Principles and Selected Applications of Item Response Theory," in Robert Linn (ed.), Educational Measurement, (New York, NY: Macmillan, 1989), pp. 147-200.

4. Koffler, op. cit., footnote 1, p. 128.

5. Ibid.; more precisely they're counted right a proportion c of the time; c is described in this paper's appendix.

6. Ibid., p. 126; Ina V. S. Mullis et al., The State of Mathematics Achievement, Executive Summary, (Washington, DC: U.S. Department of Education, National Center for Education Statistics, June 1991), p. 6.

7. e.g.: Hambleton, op.cit., footnote 3

8. Special tabulations provided by Steven Ferrara, Maryland Department of Education, Mic Lang, Louisiana Department of Education, and Joy Frechtling, Montgomery County Public Schools; the author is indebted to Arnold Packer, U.S. Department of Labor, for suggesting this approach.

9. "Spiral Math Session Script", (Princeton, NJ: Educational Testing Service, n. d.)

10. "1990 Assessment Math Public Release", (Princeton, NJ: Educational Testing Service, n. d.)

11. Ina V. S. Mullis et al., The State of Mathematics Achievement, (Washington, DC: U.S. Department of Education, National Center for Education Statistics, June 1991), p. 122; and The State of Mathematics Achievement in the District of Columbia, (Washington, DC: U.S. Department of Education, National Center for Education Statistics, June 1991), similar books appeared for each participating state and territory.

12. Steven Ferrara, Maryland Department of Education, personal communication, July 31, 1991.

13. Steven F. Ferrara and Stephen J. Thornton, "Using NAEP for Interstate Comparisons," Educational Evaluation and Policy Analysis, 10:3, Fall 1988, pp. 200-11.

14. Mathematics Objectives, 1990 Assessment, (Princeton, NJ: Educational Testing Service, November

1988)

15. e.g.: Paul Burke, "Math that Adults Need," Journal of College Admissions, Summer 1990, pp. 15-17; Colman McCarthy, "Who Needs Algebra?," Washington Post, April 20, 1991, p. A21; William Raspberry, "No. 1 in Math - the Wrong Goal," Washington Post, March 26, 1990, p. A11; William Raspberry, "I Need Math but Not Sine Curves," Washington Post, April 19, 1989, p. A19.

16. Objectives for Career & Occupational Development (Denver, CO: Education Commission of the States, 1977).

17. Arthur N. Applebee et al., Learning to Write in Our Nation's Schools, (Princeton, NJ: Educational Testing Service, June 1990).

18. Jay Simmons, "Portfolios as Large-scale Assessment," Language Arts, 67:3, March 1990, pp. 262-8.

19. See William E. Brock et al., What Work Requires of Schools, (Washington, DC: U.S. Department of Labor, Secretary's Commission on Achieving Necessary Skills, June 1991).

20. Mathematics Objectives, op. cit., footnote 14; the description of reviewing problems is in Mullis, op. cit., footnote 11, pp. 461-82 and Koffler, op. cit., footnote 1, pp. 212, 265-79; also see Robert A. Forsyth, "The NAEP Proficiency Scales: Do They Yield Valid Criterion-referenced Interpretations?," Iowa Testing Programs Occasional Papers, Number 35, May 1990.

21. Walt Haney and Laurie Scott, "Talking with Children about Tests, A Pilot Study of Test Item Ambiguity," (Cambridge MA: Huron Institute, August 1980).

22. Mullis, op. cit., footnote 11, p. 6.

23. Ibid., p. 56.

24. Ibid., pp. 466-7.

25. "NAEP 1990 National and Trial State Assessments in Mathematics," OERI Bulletin, Summer 1991, pp. 1-2.

26. Mullis, op. cit., footnote 11, p. 6.

27. Ibid, p. 56.

28. W. M. Yen, "The Extent, Causes and Importance of Context Effects on Item Parameters for Two Latent Trait Models," Journal of Educational Measurement, 17, 1980, pp. 297-311.

29. Koffler, op. cit., footnote 1, p. 142.

30. F. M. Lord, "Practical Applications of Item Characteristic Curve Theory," Journal of Educational Measurement, 14 pp. 117-38, cited in Hambleton, op. cit., footnote 3.

31. Arthur N. Applebee et al., The Writing Report Card, (Princeton, NJ: Educational Testing Service, January 1990); and Applebee, op. cit., footnote 17; emphasis added.

32. Lynn Jenkins, Educational Testing Service, personal communication, July 1990.

33. e.g.: Council for Open Discussion, Senate Guidebook, How to Get Laws You Can Live with, 1991

34. Archie E. Lapointe et al., A World of Differences, an International Assessment of Mathematics and Science, (Princeton, NJ: Educational Testing Service, January 1989).

35. John A. Dossey et al., The Mathematics Report Card, (Princeton, NJ: Educational Testing Service, June 1988).

36. Mullis, op. cit., footnote 11, pp. 55, 58, 69.

37. e.g.: Lauro F. Cavazos, "But Math IS for Everyone," San Diego Union, April 17, 1989.

38. The press reports are listed below. These include 17 articles on the 1991 math report. Three are by the same author in different papers, making similar points, so I count them as one. One of the articles was in a magazine, Newsweek, but the paper refers to them all as newspaper articles for simplicity. The articles on the 1991 NAEP report were selected by the Office of Technology Assessment, though I added the Wall Street Journal and Boston Globe articles, since other reporters referred to them:

"A Failing Grade for U. S. Kids," Atlanta Constitution February 1, 1989, pp. 1A, 4A, from staff and wire reports

Betsy White, "Reaction to Math Test Divided," Atlanta Journal, June 7, 1991, pp. E1, E4

"Americans Not so Dumb after All: Many Adults Smarter than Japanese," Atlanta Journal, February 17, 1991, p. A19

Kathy Lally, "National Student Performance on SATs Stagnant despite Reforms," Baltimore Sun, May 3, 1990, pp. 1A, 10A

Vicki Voskuil, "U. S. Math: N. D.'s kids rate No. 1," Bismarck Tribune, June 7, 1991, pp. 1A, back page

Vicki Voskuil, "Bismarck Kids are ~~#1~~ (about Average)," Bismarck Tribune, April 15, 1990, pp.

Vicki Voskuil, personal communication, July 30, 1991

Muriel Cohen, "Report Gives US Students Low Math Score," Boston Globe, June 6, 1991, p. 19

Muriel Cohen, "U. S. Pupils Fare Poorly in Math, Science Tests," Boston Globe, February 1, 1989, pp. 1, 8

Muriel Cohen, personal communication, July 30, 1991

Mary Ann Roser, "Math Skills Ring 'Alarm Bell,'" Charlotte Observer, June 7, 1991, p. 1A

Paige Williams, " Cost of Meal Stumps Pupils," Charlotte Observer, June 7, 1991, pp. 1B, 5B

Jane Norman, "Iowa Third in Math Test in U. S., but Not 'Cutting it,'" Des Moines Register, June 7, 1991, pp. 1A, 8A

Jane Norman, personal communication, July 31, 1991

Robert Rothman, "NAEP Panel Sets Three Standards for '90 math Test," Education Week, May 22, 1991, pp. 1, 25

Robert Rothman, "States Take Stock of Math Programs in Wake of NAEP Results," Education Week, June 19, 1991, pp. 1, 16

Robert Rothman, "First State-Level Assessment Finds Wide Variations," Education Week, June 12, 1991, pp. 1, 23

Robert Rothman, personal communication, July 26, 1991

Paul Richter, "State near Bottom in Math Ranking," Los Angeles Times, June 7, 1991, pp. A3, A34

Mary Ann Roser, "Test Results Add up to This: American Students Can't Count," Miami Herald, June 7, 1991, pp. 1A, 17A

Melinda Beck et al., "New York meets Lake Wobegon," Newsweek, July 8, 1991, pp. 48-9

Barbara Kantrowitz and Pat Wingert, "A Dismal Report Card," Newsweek, June 17, 1991, pp. 64-7

Karen De Witt, "Eighth Graders' Math Survey Shows No State Is 'Cutting It,'" New York Times, June 7, 1991, pp. A1, D16

"U. S. Students Place Low on Math and Science Tests," New York Times, February 1, 1989, p. B6, story from AP

Albert Shanker, "Students Flunk Again," New York Times, January 14, 1990, p. E7

Mary Ann Roser, "Report Charts U. S. Students' Shortcomings in Math," Philadelphia Inquirer, June 7, 1991, p. 4A

Rob Walker, "State Pupils in Middle of Pack on U. S. Test Scores," Richmond Times-Dispatch, June 7, 1991, p. C1

Nanette Asimov, "Lower S. F. School Test Scores Blamed on Money Problems," San Francisco Chronicle, June 5, 1991, pp. A13-14

Nanette Asimov, personal communication, July 30, 1991

Ann Blackman, Time, personal communication, July 26, 1991

Pat Ordovensky, "U. S. Math Skills Ring 'Alarm,'" USA Today, June 7-9, 1991, pp. 1A, 10A

Dennis Kelly USA Today, personal communication, July 31, 1991

Gary Putka and Hilary Stout, "Tradition Cited as Factor in Math Scores," Wall Street Journal, June 7, 1991, p. B1

"U. S. Students Placed," Wall Street Journal, February 1, 1989, p.A1, 1 inch summary

Carol Innerst, "D. C. at Back of Class in Math," Washington Times, June 7, 1991, pp. A1, A9

Carrie Dowling, "D. C. Grade Schools Show Decline on Taking 'Higher-standard' Test," Washington Times, July 23, 1987, p. B6

Lynda Richardson, "D. C. School Superintendent Unveils New Math Program," Washington Post, July 11, 1991, p. C1

Robert J. Samuelson, "The School Reform Fraud," Washington Post, June 19, 1991, p. A19

Kenneth J. Cooper, "U. S. Youth Fail Math Test," Washington Post, June 7, 1991, pp. A1, A16

Lynda Richardson, "D. C. Scores Low in National Math Test," Washington Post, June 7, 1991, p. A15

Paul Burke, "U. S. Students: the Myth of Massive Failure," Washington Post, August 28, 1990, p. A17

Steve Twomey, "SAT Scores Fall in Va., Md.; Private Schools Boost D. C. Average," Washington Post, August 28, 1990, pp. D1, D4

Kenneth J. Cooper, "SAT Gains in D. C., Md. the Highest in the Nation," Washington Post, May 3, 1990, pp. A1, A15

Kenneth J. Cooper, "Tests of U. S. Students Show Little Progress," Washington Post, January 10, 1990, pp. A1, A5

Kenneth Cooper, personal communication, July 26, 1991

39. Lapointe, op. cit., footnote 3, cited in Cohen, op. cit., footnote 38.

40. "U. S. Students Place Low ...," op. cit., footnote 38.

41. Mullis, op. cit., footnote 6, cited in Walker, op. cit., footnote 38.

42. Kenneth Cooper, Washington Post, July 26, 1991, and Dennis Kelly, USA Today, July 31, 1991, personal communications.

43. Putka and Stout, op. cit., footnote 38.

44. Ann Blackman, Time, personal communication, July 26, 1991.

45. Voskuil, op. cit., footnote 38.

46. "Americans Not So Dumb ...," op. cit., footnote 38.

47. Virtually all the personal communications mentioned in footnote 38 raised this issue.

48. Press packet prepared by U. S. Department of Education, National Center for Education Statistics, June 6, 1991.

49. Vicki Voskuil, Bismarck Tribune, personal communication, July 30, 1991.

50. Ina Mullis, Educational Testing Service, Aug. 1, 1991 and Eugene Owen, U.S. Department of Education, National Center for Education Statistics, August 1, 1991, personal communications.

51. Rpbert Rothman, Education Week, July 26, 1991, Kenneth Cooper, Washington Post, July 26, 1991, and Dennis Kelly, USA Today, July 31, 1991, personal communications.

52. Ramsay W. Selden, "Charting and Adjusting Test Scores," Education Week, September 13, 1989, pp. 32, 27.

53. Mullis, op. cit., footnote 11, p. 7.

54. Ibid., p. 479.

55. US Department of Labor, Bureau of Labor Statistics, "The Employment Situation: June 1991," July 3, 1991.

56. US Department of Health and Human Services, "HHS News," May 1, 1991.

57. David C. Hammack et al., The U. S. History Report Card, (Princeton, NJ: Educational Testing

Service, April 1990).

58. Mullis, op. cit., footnote 6, p. 1.

59. Special tabulations for 1990 from the U. S. Department of Labor, Division of Local Area Unemployment Statistics.

60. Mullis, op. cit., footnote 6, p. 1.

61. Special tabulations for 1989 from the U. S. Department of Labor, Division of Foreign Labor Statistics and Trade; data are at purchasing power parities.

62. Mullis, op. cit., footnote 6, p. 1.

63. Eugene Owen, U.S. Department of Education, National Center for Education Statistics, August 1, 1991, personal communication.

64. Koffler, op. cit., footnote 1, p. 18.

65. Mullis, op. cit., footnote 6, pp. 437-9.

66. Dossey, op. cit., footnote 35, p. 126.

67. Jane Norman, Des Moines Register, personal communication, July 31, 1991; for more on limitations, see: Iris C. Rotberg, "I Never Promised You First Place," Phi Delta Kappan, December 1990, pp. 296-303; Norman Bradburn et al., "A Rejoinder to 'I Never Promised You First Place,'" Phi Delta Kappan, June 1991, pp. 774-777; Iris C. Rotberg, "How Did All Those Dumb Kids Make All Those Smart Bombs?," Phi Delta Kappan, June 1991, pp. 778-81; John B. Carroll, "The National Assessments in Reading," Phi Delta Kappan, February 1987, pp. 424-30.

68. Haney, op. cit., footnote 21.

69. e.g.: Joint Matriculation Board Examinations Council, "GCE Examiners' Reports," 1987; West African Examinations Council, Regulations and Syllabuses for the Joint Examinations for the School Certificate and General Certificate of Education (Ordinary Level) and for the General Certificate of Education (Advanced Level), 1987-88, n. d.

70. Facts about the questions are shown in Koffler, op. cit., footnote 1, pp. 22, 175, 247-52 and Mullis, op. cit., footnote 11, pp. 60-78, 466-82, 506-8. It would be helpful to the reader and an aid in finding context effects if this information were printed together, and if blocks 3, 7 and 9, which have been released, were printed in their entirety, along with the introductory explanations

for the student and the script for the test administrator. It is also very important to publish the percent of students who choose each wrong answer or skip a problem, so we can begin to know what students' misconceptions are and design teaching to address them. This information is not published at all now.

71. Ibid., drawn by the author.

72. Koffler, op. cit., footnote 1, p. 132.

73. See footnote 70; drawn by author.

74. John Mazzeo and Kentaro Yamamoto, "The 1990 NAEP Mathematics Scale," presented at the annual meeting of the National Council on Measurement in Education, Chicago, April 1991; Koffler, op. cit., footnote 1, p. 136.

75. Koffler, op. cit., footnote 1, p. 136.

76. Mullis, op. cit., footnote 6, p. 6.

77. Hartvig Dahl, Word Frequencies of Spoken American English, (Essex, CT.: Verbatim, 1979)

78. White, op. cit., footnote 38.

79. See footnote 70; drawn by author.

80. "Mathematics Objectives," op. cit., footnote 14, p. 56.

81. Koffler, op. cit., footnote 1, p. 126.

82. See footnote 70; drawn by author.

83. Robert J. Mislevy, "Bayes Modal Estimation in Item Response Models," Psychometrika, 51:2, June 1986, pp. 177-95; also see: Hariharan Swaminathan and Janice A. Gifford, "Estimation of Parameters in the Three-Parameter Latent Trait Model," in D. Weiss (ed.), New Horizons in Testing, Academic Press, 1983, pp. 13-30;and Hariharan Swaminathan and Janice A. Gifford, "Bayesian Estimation in the Three-parameter Logistic Model," in Psychometrika, 51, pp. 589-601

84. Koffler, op. cit., footnote 1, pp. 145, 149.

85. Ibid., p. 127.

86. Ibid., pp. 131-2

87. Ibid., pp. 129-33; also see Eugene Johnson and Robert Mislevy, "Accounting for Measurement Errors in Estimation ... Reading Proficiency," Educational Testing Service, Nov. 6, 1990.

88. Robert J. Mislevy et al., "Estimating Population Characteristics from Sparse Matrix Samples of Item Responses," forthcoming.

89. Hambleton, op. cit., footnote 3; Fumiko Samejima, "A Use of the Information Function in Tailored Testing," Applied Psychological Measurement, 1:2, Spring 1977, pp. 233-47.

90. See footnote 70.

91. Information scale is based on a conventional distribution with mean 0, standard deviation 1, for ease of comparison with other articles.

92. See footnote 70; drawn by author.

93. Ibid.

94. Ibid.

95. Ibid.